クラスター分析におけるクラスター数自動決定法の比較

志津綾香* 松田眞一†
E-Mail: matsu@nanzan-u.ac.jp

本論文ではクラスター分析における種々のクラスター数自動決定法を概観し、その一部について比較検討を行う。既存のプログラムおよび自ら作成したプログラムを用いていくつかの設定のもとでモンテカルロシミュレーションで比較する。比較した中ではいくつかの方法が優秀であったが、状況によって優劣が変化することが分かった。その中でも特に有力なのは Jain-Dubes(k-means) 法と Upper Tail(ward) 法であるが万能ではないので、現実的には他の有力な方法とも組み合わせながら判断するのがよいと考えられる。

1 はじめに

多くの統計解析の調査研究では、類似度の評価を頼りに様々な分析を行う。その中でもクラスター分析は、類似度を評価するための基本的な方法であり、誰でも行うことができる数量的方法である。しかし、クラスター分析は一般的に確率的評価と結びつけることがないため、分析で最も大きな意味を持つクラスター数に関して解析者自身で適当に決定するのが普通である。そのための判断基準はデンドログラムを視覚的に眺めるなど曖昧なことが多い。

本論文では、クラスター数自動決定法にどのようなものがあるかを概観し、その一部についてモンテカルロシミュレーションで比較し、評価を与えることを目的とする。

2 記号

本論文では p 次元空間にある大きさ n の標本を m 個のクラスターに分割する問題を考える。

距離（非類似度）を用いる場合は 2 つの標本点 $\boldsymbol{x}_1, \boldsymbol{x}_2$ に対して $D(\boldsymbol{x}_1, \boldsymbol{x}_2)$ で表し、標本点とクラスター、クラスターとクラスターの距離も同じ D という記号を用いることとする。

各クラスターは C_i で表し、その重心点となる標本平均ベクトルは $\boldsymbol{c}_i$ で表わす。また、n_i はクラスター C_i 内の標本点の数とする。

3 クラスター分析

3.1 クラスター分析とは

クラスター分析とは、2 つ以上のデータがあるとき、類似度や距離 (非類似度) を手がかりに、データをいくつかのグループ（それをクラスターと呼ぶ）に分類する方法である。

クラスター分析は大別すると、階層的方法と非階層的方法の 2 つの計算方法がある。

*南山大学大学院数理情報研究科数理情報専攻
†南山大学情報理工学部情報システム数理学科

階層的方法で主に用いられる方法は、最短距離法、最長距離法、群平均法、重心法、メジアン法、ward 法であり、クラスターを形成する際のクラスター生成法が異なる。その基準となる距離には、ユークリッド距離や標準化ユークリッド距離およびマハラノビス距離等がある。(一般的には(標準化)ユークリッド距離が使用されることが多い。)すなわち、階層的方法は距離とクラスター生成法の組み合わせで方法が決定する。非階層的方法には、k-means 法、超体積法等がある。非階層的方法の場合、ある評価関数を基準にクラスターを生成する。(菅 [7], 渡辺ら [17], 神嶌 [6], 田中・脇本 [16] 参照)

本論文では階層的方法ではユークリッド距離による最長距離法と ward 法、非階層的方法では k-means 法を比較の対象として扱う。

4 階層的方法

階層的方法は、まず各標本点を 1 つのクラスターとして、最も距離の近い標本点から順に合併させ、新たなクラスターを形成していく方法である。結果はデンドログラムと呼ばれる樹形図で表示させることができ、似ているものから順に枝の合流でクラスター生成が示されている。そして、一般的に、クラスター数はそのデンドログラムを見て決定される。クラスター分析において、どのようなクラスターが形成されるかはクラスター生成法に依るところが大きい。

以下では本論文で取り上げるクラスター生成法のみを説明する。

4.1 最長距離法

最長距離法とは、各クラスター間の距離における最長距離を、クラスター間の距離とする方法である。この方法は、1 つのクラスターが極端に大きくなるのを抑えられ、大きさのそろったクラスターを得ることができる。

$$D(C_1, C_2) = \max_{\boldsymbol{x}_1 \in C_1, \boldsymbol{x}_2 \in C_2} D(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

4.2 ward 法

ward 法は、クラスターとしてサンプルをまとめるときに生じる、各サンプルの情報の損失量の増加分をクラスターの距離とする方法である。すべてのクラスター内の偏差平方和の和をできるだけ小さくするように組み合わせていくので、比較的まとまりのあるクラスターがいくつか得られる。

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

ただし、$E(C_i) = \sum_{\boldsymbol{x} \in C_i} (D(\boldsymbol{x}, \boldsymbol{c}_i))^2$ である。

5　非階層的方法

非階層的方法は、あらかじめいくつのクラスター数にするかを決めておき、その数に従って標本点を振り分けていくというものである。決められたクラスター数に対して、できるだけクラスター間の距離は大きく、各クラスターの標本点間の距離は小さくなるように新たな標本点を振り分けていく方法であるので、クラスター数を変えて分析を行うと生成されたクラスター間に包含関係がないことが多い。非階層的方法は、計算量が膨大なため、処理時間が長くなるのが欠点である。

5.1　k-means 法

k-means 法は、あらかじめクラスター数を決めておき、各標本点を振り分けていく方法である。クラスターに含まれる各標本点とそのクラスターの重心点の距離が、他のどのクラスターの重心点よりも小さくなるように求める。各クラスター間の距離は大きく、クラスター内の距離は小さくなるように分割されている。つまり、計算方法としては、分割のよさの評価関数を定め、その評価関数を最適にする分割を探し当てることになる。

評価関数は、次のように与えられる。

$$\sum_{i=1}^{m} \sum_{\boldsymbol{x} \in C_i} (D(\boldsymbol{x}, \boldsymbol{c}_i))^2$$

k-means 法は、クラスターの重心点をクラスターの代表点とし、評価関数を、最小化する m 個のクラスターに分割する。最適解は、対象点のクラスターへの割り当てと、代表点の再計算を交互に繰り返し行って探す。この方法は、局所最適解しか求められないため、ランダムに初期値を変更して、評価関数を最小する結果を選択するのが一般的である。

6　クラスター数決定法

クラスター数が未知の場合に、最適クラスター数を求める方法には、大きく分けて以下の 4 つの方法がある。(石岡 [2] 参照)

1. クラスター数の変えてクラスター分析を何通りか行い、情報量基準などの適当な基準を用いて最適なクラスター数を求める方法
2. 最小体積楕円体推定量を用いる方法 (Jolion et al.[5] 参照)
3. 最適解と思われるクラスター数より少し多い数のクラスター分割から始めて、近いクラスター同士を併合したり、外れ値などの不要なクラスターを抹消することにより、適当な数のクラスターを決定する方法 (Krishnapuram and Freg[8] 参照)
4. x-means 法と呼ばれ、最初にある十分小さい数のクラスターに k-means 法で分類した後、各クラスターに対して同様に k-means 法による 2 分割を、その分割が適当でないと判断されるまで繰り返す方法 (Pelleg and Moore[14] 参照)

1. は最も単純な方法であるが、一般に情報量基準などの評価関数がクラスター数に対して単峰とならないことが多く、最適なクラスター数の決定が局所最適解に陥る可能性がある。

ここに属する古典的幾何学法は、群の数である m と分割の判断基準の値をプロットし、人間の目で見て判断するというものであるため、自動化は行えない。その他の方法は、次節で紹介する。

2. は各クラスターは楕円体で分割されるという仮定のもとで、全体から始め Kolmogrov-Smirnov 検定を用いて、単一の楕円体とみなすことのできるクラスターを順次決定し、取り除いていくものである。しかし、この方法は、検定を行う際の有意水準の決定が難しいこと、またデータの外乱に極めて弱いことが知られている。(Nasraoui et al.[12] 参照)

3. は全てのデータを排他的に分類するのではなく、外れ値とみなされるデータをクラスターから除外することに重きを置いている。そのため、一般的なクラスター数自動決定に向いているわけではない。

4. は最終的にでき上がるクラスターの数が不定であることから、x-means 法と呼ばれている。この x-means 法は、石岡 [2] によって改良が提案され、そのアルゴリズムが R 上の関数で実装されている。

本論文では、1. の一部と 4. の方法を利用する。

7　クラスター数自動決定法

本論文で用いたクラスター数自動決定法と、前述した主なクラスター数決定法を簡単に紹介する。

7.1　Jain-Dubes 法

Jain-Dubes 法（以下では JD 法と略す）は Jain and Dubes[4](Ngo et al.[13] 参照) による以下の数式を用いる決定法である。

$$p(m) = \frac{1}{m}\sum_{i=1}^{m} \max_{1\leq j\leq m}\left\{\frac{\eta_i + \eta_j}{\xi_{ij}}\right\}$$

ここで

$$\eta_j = \frac{1}{n_j}\sum_{i=1}^{n_j} D(\boldsymbol{F}_i^{(j)}, \boldsymbol{c}_j)$$

$$\xi_{ij} = D(\boldsymbol{c}_i, \boldsymbol{c}_j)$$

であり、$\boldsymbol{F}_i^{(j)}$ はクラスタ C_j 内の i 番目のベクトルとする。

最適なクラスター数は、$p(m)$ の値をある範囲内で最小にするような m となる。しかし、上記の基準の定義上、全てのクラスターが単独サンプルになるとき、$p(m)$ が 0 になり一番最適となってしまうので、ある範囲の決定が重要である。本論文では、適当な目安として、

スタージェスの公式 $1+\log_2 n$ を用いた。すなわち、m の範囲は $(1 \leq m \leq 1+\log_2 n)$ とする。(志津 [15] 参照)

7.2 x-means 法

x-means 法は、k-means 法の拡張である。あらかじめクラスター数を決めておかなければならない k-means 法とは違い、最適なクラスター数を推測することができる。x-means 法の考え方は、$m=2$ で再帰的に k-means 法を実行していくというもので、クラスタの分割前と分割後で BIC 値（ベイズ情報量規準）を比較し、値が改善しなくなるまで分割を続ける。

つまり、分割前のベイズ情報量を BIC、分割後のベイズ情報量を BIC′ とし、

- BIC>BIC′ ならば 2 分割する
- BIC≤BIC′ ならば 2 分割しない

を全ての場合で 2 分割できなくなるまで繰り返すことにより最適なクラスター数を決定する。

BIC 値は以下のように定義される。p 変量正規分布を

$$f(\boldsymbol{\theta}_i;\boldsymbol{x}) = (2\pi)^{-p/2}|V_i|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t V_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

と仮定すると、

$$BIC = -2\log L(\hat{\boldsymbol{\theta}}_i;\boldsymbol{x}_i \in C_i) + q\log n_i$$

ここで、$\hat{\boldsymbol{\theta}}_i$=[$\hat{\boldsymbol{\mu}}_i$,$\hat{V}_i$] は、$p$ 変量正規分布の最尤推定値とする。$\boldsymbol{\mu}_i$ は p 次の平均値ベクトル、V_i は $p\times p$ の分散共分散行列である。q はパラメータ空間の次元数で、V_i の共分散を無視すれば（0 とおくことが可能ならば）q=$2p$ である。共分散を無視しなければ $q=p(p+3)/2$ である。L は尤度関数で $L(\cdot)=\prod f(\cdot)$ である。

2 分割したモデルにおける BIC′ は以下のようになる。

$$BIC' = -2\log L(\hat{\boldsymbol{\theta}}'_i;x_i \in C_i) + q'\log n_i$$

ここで、$\hat{\boldsymbol{\theta}}'_i$=[$\hat{\boldsymbol{\theta}}^1_i$,$\hat{\boldsymbol{\theta}}^2_i$] は、2 分割されたそれぞれの p 変量正規分布の最尤推定値とする。共分散を無視すれば（0 とおくことが可能ならば)、各 p に対し平均と分散の 2 つのパラメータが存在するので、パラメータ空間の次元は $q'=2\times 2p=4p$ となる。共分散を無視しなければ $q'=2p=p(p+3)$ である。

本論文のシミュレーションでは、クラスター数をカウントするために少しだけプログラムを変更したが、x-means 法の計算自体は石岡 [2] によって改良された x-means 法をそのまま用いた。プログラムは R および Fortran(g77) で実装されており、そのソースコードは石岡 [3] より入手することができる。x-means 法の詳しいアルゴリズムは、石岡 [2] を参照のこと。

7.3 Upper Tail 法

Upper Tail 法（以下では Tail 法と略す）は、Mojena[10] によって提案され、階層的方法における重要なクラスター数決定法である。統計的な停止規則を用いてクラスター数を求める。その方法は、大きさ n の標本に対してクラスターを生成するための基準値 α が $n-1$ 個あるのを利用する。ここでは基準に距離のみを考えるのですべてが一つになる距離 α_1 から降順に α_{n-1} までの基準値がある。この α の分布の平均と標準偏差を計算することによって、有意な α を導くことでクラスター数を決定する方法である。停止規則は、$j=1$ から始めて条件

$$\alpha_j \leq \bar{\alpha} + ks_\alpha$$

を満たすまで j を増加させることである。停止した j が最適なクラスター数となる。$\bar{\alpha}$ と s_α はそれぞれ α の分布の平均と不偏分散の平方根をとったものである。この方法は最短距離法、最長距離法、群平均法、ward 法で利用可能である。

k の値については、Mojena[10] では、2~4 の数を使っている。その時に用いたデータは、データの総数が 60~120 のものを、2~4 に分割したものである。本論文では、さらに 1 群のデータ数が多い場合など、k の値について掘り下げていきたい。

7.3.1 Upper Tail 法の改良

本研究では、Tail 法の改良を提案する。Tail 法の停止規則を計算する時、従来の方法ではデータが正規分布に従うよう求めているが、カイ 2 乗分布に従うように計算する。一般的に多変量正規分布に従う 2 点間の距離の分布はカイ 2 乗分布に従うからである。具体的な計算方法は、以下のように α を正規化し

$$\alpha' = \Phi^{-1}(F_p(\alpha/s_\alpha \cdot p))$$

それに前節の Tail 法を用いている。ここで Φ^{-1} は正規分布の逆関数であり、F_p は自由度 p のカイ二乗分布の分布関数である。

7.4 その他の方法

その他のクラスター数自動決定法として、凸集合の推測に基づく方法（Hardy[1], Moore[11] 参照）、クラスターのための尤度比検定（Hardy[1] 参照）、移動平均品質管理規則（Mojena[10] 参照）、Wolf[18] のテスト、Marriot[9] のテストがあるが、詳細は割愛する。

8 シミュレーション

8.1 シミュレーションに用いるデータ

今回シミュレーションで用いるデータの形は、主に次の 4 種類である。

8.1.1　パターン 1

2 次元の正規乱数を作り、それらを 3 つ用意する。その各データの平均を 1 つの正三角形となるように置く。データ数、相関、分散を変化させる。例として、図 1 に、軸となるデータを載せておく。このとき、1 つの群の乱数データの数が 100 個、各データの平均はそれぞれ (0,0),(3,0),(1.5,1.5× $\sqrt{3}$)、分散は (1,1)、相関係数は 0 である。

8.1.2　パターン 2

2 次元の正規乱数を作り、それらを 3 つ用意する。その各データの平均を均等に直線上に並ぶように置く。データ数、相関、分散を変化させる。

データの例は図 2 のようになる。このとき、1 つの群の乱数データの数が 100 個、各データの平均はそれぞれ (0,0),(1.5,0),(3,0)、分散は (1,1)、相関係数は 0.5 である。

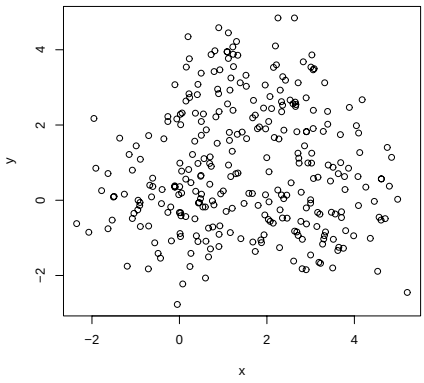

図 1: パターン 1

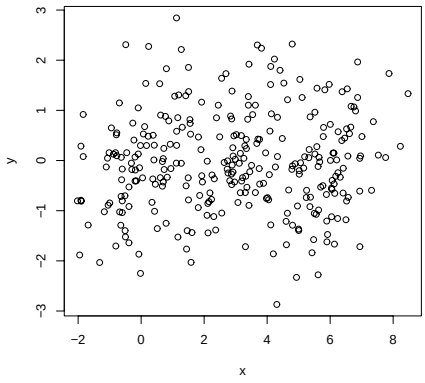

図 2: パターン 2

8.1.3　パターン 3

ある分布に従う 2 次元の乱数を作り、それらを 3 つ用意する。その各データの平均を 1 つの正三角形となるように置く。パターン 1 の分布を変化させたものである。今回は t 分布、ラプラス分布、対数正規分布を用いる。分布によって、平均の位置、自由度、標準偏差等を変化させる。データの例は図 3 のようになる。これは、対数正規分布に従い、1 つの群の乱数データの数が 100 個、各群のデータの平均がそれぞれ (1,1),(4,1),(2.5,1+1.5×$\sqrt{3}$) となるように配置したときのプロット図である。今回、対数正規分布に従う場合の全ての群の標準偏差を揃えている。図 3 では 0.5 である。

8.1.4　パターン 4

2 次元の正規乱数を作り、それらを 5 つ用意する。その 5 つの乱数を、十字型になるように並べ、分散を変化させる。例として、軸となるデータを図 4 に載せておく。このとき、1 つ

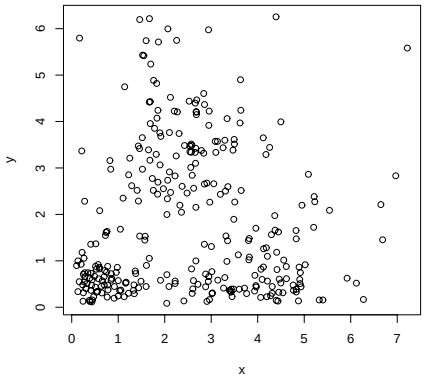

図 3: パターン 3

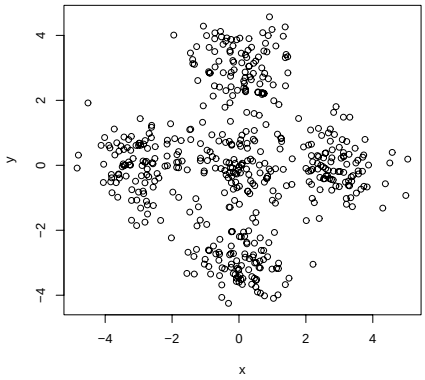

図 4: パターン 4

の群の乱数データの数が 50 個、各データの平均はそれぞれ (0,0),(3,0),(0,3),(−3,0),(0,−3)、分散は (0.5,0.5)、相関係数は 0 である。

8.2 比較するクラスター数自動決定法

シミュレーションを行うために、ある分布に従う 2 次元の乱数をつくり、それらを適切な形にして階層的手法（最長距離法と ward 法）と非階層的手法（k-means 法）の場合で、クラスター数自動決定法の計算をさせるプログラムを作成した。

シミュレーションで用いたクラスター数自動決定法は、JD 法, Tail 法, x-means 法である。JD 法では階層的手法、非階層的手法の両方が利用できるので、最長距離法、ward 法、k-means 法の 3 方法で実験する。Tail 法は階層的手法にのみ適用可能なので、最長距離法と ward 法の 2 方法のみを用いる。また今回は 7.3.1 節のようなカイ 2 乗分布に基づく Tail 法の改良も試みた。よって、Tail 法では従来の正規分布に基づく 2 方法の場合とカイ 2 乗分布に基づく 2 方法、全部で 4 方法のシミュレーションを行う。

<table>
<tr><td rowspan="3">JD 法</td><td rowspan="3"></td><td>最長距離法</td></tr>
<tr><td>ward 法</td></tr>
<tr><td>k-means 法</td></tr>
<tr><td>x-means 法</td><td></td><td></td></tr>
<tr><td rowspan="4">Tail 法</td><td rowspan="2">正規分布</td><td>最長距離法</td></tr>
<tr><td>ward 法</td></tr>
<tr><td rowspan="2">カイ 2 乗分布</td><td>最長距離法</td></tr>
<tr><td>ward 法</td></tr>
</table>

表 1: シミュレーションで使う 8 方法

この時、計算の繰り返し数、1 つの乱数におけるデータ数、各平均、各分散、相関係数を

入力する。結果は、3 群、5 群に分けたデータであるので、最適なクラスター数がそれぞれ 3、5 となった時の数をカウントし、表示させる。

8.3　シミュレーション条件の決定

8.3.1　繰り返し数と標本の大きさについて

志津 [15] で繰り返し数 200 回程で、1000 回繰り返した時とほぼ変わりない結果が得られることが分かったので、本論文のシミュレーションでも繰り返し数を 200 回とすることにした。

また、今回のシミュレーションでは、コンピュータの計算にかかる時間の都合上、1 つの群の標本の大きさは主に 100 個前後とする。

以上のような条件で、更に分散や相関係数、分布を変化させたときの、各クラスター数自動決定法における最適なクラスター数の変化の動きなどを調べていく。

8.3.2　シミュレーションの設定

以下の設定でシミュレーションを行う。

A. パターン 1,2,4 において分散、相関を変化させた時

B. パターン 3 の t 分布, ラプラス分布において各平均を変化させた時

C. パターン 3 の対数正規分布においての各標準偏差を変化させた時

D. パターン 1 において、3 つの乱数全てのデータ数が違う場合で分散を変化させた時

E. パターン 1 において、3 つの乱数全ての分散が違う場合で平均を変化させた時

以上の条件において、JD 法を用いた 3 方法と Tail 法を用いた 4 方法と x-means 法でシミュレーションを行い、考察する。また、Tail 法についてのみ k の値を変化させた時も実験し、Mojena[10] でも明記されていない k の値について、考察する。

9　シミュレーション結果

9.1　正規分布で標本の大きさが揃っている場合

9.1.1　設定 A でパターン 1 の場合

パターン 1 において、分散を 0.5 から 2 まで変化させた時の、JD 法と Tail 法の当てはまり具合は図 5 のようになった。データ数は 100×3 個、相関 0 で固定である。

JD 法を用いた場合で一番良い結果となったのは、k-means 法である。Tail のカイ 2 乗分布を用いた時の最長距離法と ward 法が、分散が強くなった時にも耐えて良い結果を出している。Tail を用いた最長距離法の場合、ほとんど正解できていないので、カイ 2 乗分布を用いたことでかなり改善されたといえる。

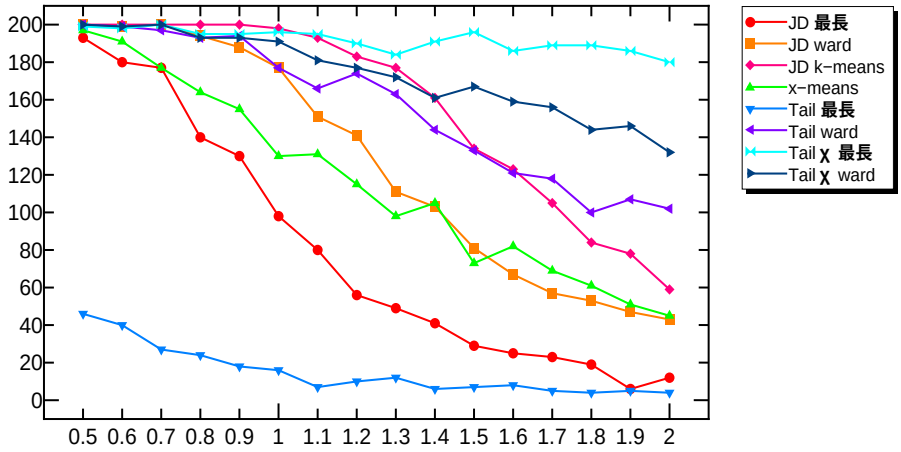

図 5: パターン 1 で分散を変化させた時

グラフは省略するが、相関を 0 から 1 まで変化させた場合も、分散を変化させた時と似たような結果になった。分散は 1 で固定である。JD の k-means 法は極端なデータでなければ安定して良い結果が得られる。しかし相関が 0.7 を超え、データの形が極端になると、Tail のカイ 2 乗分布を使った最長距離法と ward 法が強くなる。

9.1.2 設定 A でパターン 2 の場合

パターン 2 において、相関を 0 から 1 まで変化させた時の、JD 法と Tail 法の結果は図 6 のようになった。データ数は 100 × 3 個、分散 1 で固定する。相関 0.5 以上は人の目でみて 3 群と認識できるようなデータとなっている。

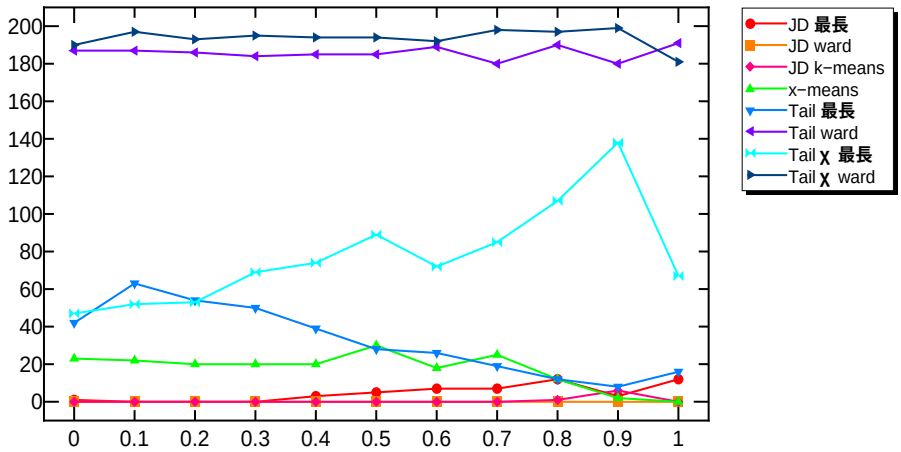

図 6: パターン 2 で相関を変化させた時

分散 1 で固定した場合、データの重なりはそこまで大きくなくても、あまり離れていな

いため、このような極端な結果になった。一方、分散を 0.5 で固定した場合、相関をいくつに変化させても、常にくっきり 3 群に分かれている結果が得られる。すなわち、常に正解する Tail の ward 法、カイ 2 乗の ward 法、5〜8 割の正解率の JD の k-means 法、x-means 法、Tail の最長距離法、カイ 2 乗の最長距離法、1 割以下しか当たらない JD の最長距離法、ward 法である。真ん中の 4 方法は、明らかにデータが分かれているような場合でないと、判別するのが難しいことが分かる。また、JD 法を用いた場合の ward 法は、三角のデータなどではそれなりに良い結果を残しているが、パターン 2 のように縦長に広がるデータの場合が、かなり苦手と言えるだろう。

なお、Tail 法のカイ 2 乗の場合は相関が 1 のときに急に下がるグラフとなったが、他の場合では急に上がる場合もある。どちらもつぶれて自由度が変わったことで反応が変わったと考えられる。しかし、相関 1 は特殊な状況なので実用上は問題ないだろう。

9.1.3 設定 A でパターン 4 の場合

パターン 4、正規分布に従う 5 群からなるデータにおいて、分散を 0.5 から 2、相関を 0 から 1 まで変化させた時の、JD 法と Tail 法の結果は図 7 のようになった。データ数は 50 × 5、分散を変化させる時は相関を 0 で固定し、相関を変化させる時は分散を 1 で固定した。

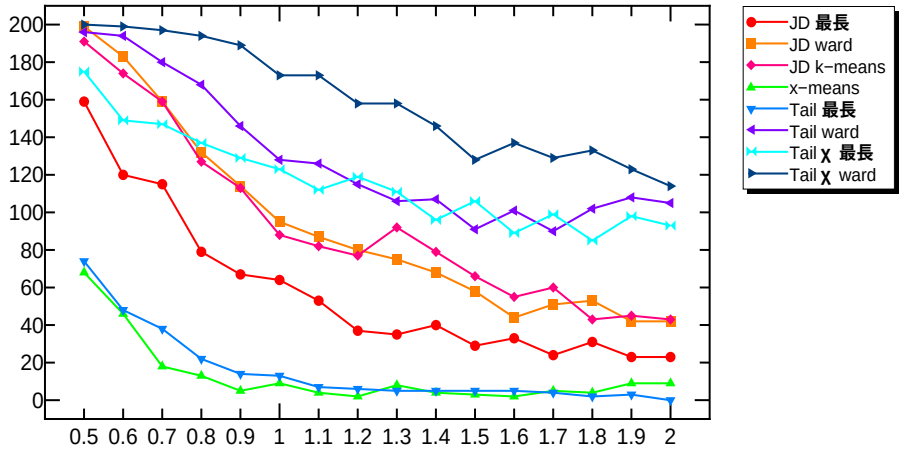

図 7: パターン 4 で分散を変化させた時

分散を変化させた場合、3 群の時とほぼ同じような結果が出た。違うのは、3 群の時に飛びぬけて良かったカイ 2 乗分布を用いた Tail の最長距離法が 5 群だと少し正解率が全体的に下がってしまった所である。

相関を変化させた時も、分散を変化させた時の結果と、順位的にはほとんど似たものになった。設定 A の 3 群ではかなり正解率の高かった JD の k-means 法も、5 群の場合になると少し精度を下げた。一貫して強いのは、Tail の ward 法、カイ 2 乗分布を用いた Tail の ward 法である。

9.2　正規分布ではない場合

9.2.1　設定 B の場合

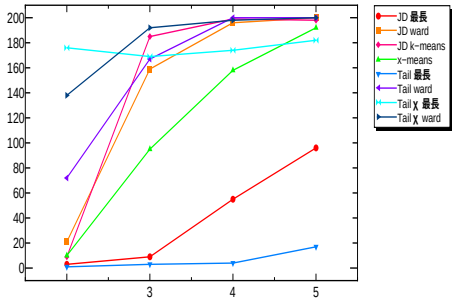

図 8: パターン 3(t 分布) で平均の距離を動かした時

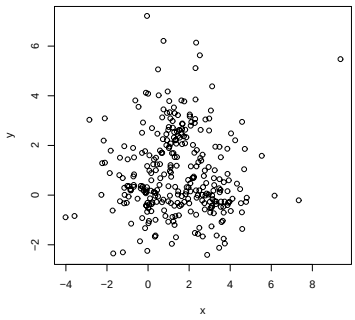

図 9: t 分布:mu=3 の時

乱数データの各群の平均間の距離 mu が 3 の時の t 分布に従うデータは図 9 のようになる。自由度は 5 とした。人の目でみて、ギリギリ 3 群と判断できるかというレベルのデータである。図 8 によると、JD と Tail の最長距離法は、$mu = 5$ での、はっきりと 3 群に分かれている場合でも結果は良くなかった。カイ 2 乗分布を用いた Tail の最長距離法、ward 法は、他の方法が苦手だった $mu = 2$ の場合もかなり高い正解率を得ており、カイ 2 乗分布を用いた場合の有利な点と言える。

ラプラス分布の結果は、t 分布の結果とほとんど同じ形になったので詳細は割愛する。また、t 分布に関してはパターン 4 と同じ 5 群の場合も検討したが、傾向に大きな変化はなかった。

9.2.2　設定 C の場合

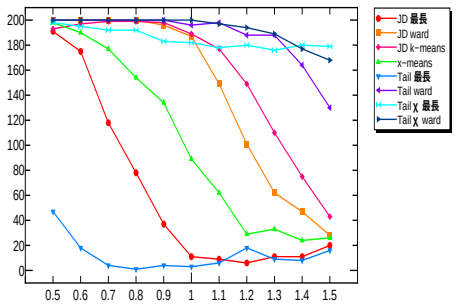

図 10: パターン 3(対数正規分布) で標準偏差を変化させた時

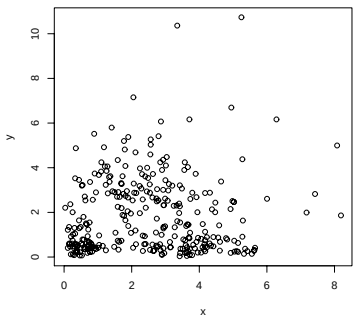

図 11: 対数正規分布:標準偏差 1.2 の時

図 11 は標準偏差 1.2 の時のデータであるが、割と高い正解率を出す JD の k-means 法や JD の ward 法の正解率が落ち始めるときのものである。図 10 によると、対数正規分布の場合も、t 分布の時同様、他の方法が正解率を下げる標準偏差 1.2 以降においても、Tail のカイ 2 乗分布を用いた最長距離法と ward 法、Tail の ward 法が圧倒的な強さを見せた。分布の対称性が崩れてもこれらの方法は強いようである。また、パターン 4 と同じ 5 群の場合も検討したが、x-means 法が少し弱くなった以外は傾向に大きな変化はなかった。

9.3 正規分布で標本の大きさが異なる場合や不等分散の場合

設定 D は、データ数を 50,100,200 で行った。結果は Tail のカイ 2 乗分布を用いた最長距離法が圧倒的な強さをみせた。しかし、これは k の値がかなり適合したからだろう。だからといって、Tail のカイ 2 乗分布を用いた ward の方は徐々に精度が下がっているので、k の値抜きにしてもカイ 2 乗分布を用いた最長距離法が、データ数が違う場合に強い事が分かる。

設定 E は、分散 1,2,4 で行った。結果は Tail のカイ 2 乗分布を用いた最長距離法が、他の方法では正解率が低い場合も、高い正解率を出した。Tail の最長距離法は、この場合もうまくいかなかったが、その他の Tail を用いた方法は基本的に各群の分散がそれぞれ違っていてもそこまで結果に影響されないと言える。

9.4 Tail 法において k の値を変化させた時

9.4.1 3 群の場合

パターン 1 において、分散 1、相関 0 でデータ数 50×3 の時とデータ数 100×3 の時、k を変化させる。

図 12, 13 から、k の値によって正解率が変わることが分かる。また、今回用いた 4 方法全てにおいて、それぞれ正解率がほぼ 100 パーセントになるピークの時があると言える。そしてそのピークは、データ数によってずれている。

9.4.2 5 群の場合

次に、5 群のデータについて k 変化時の動きをみてみる。パターン 4 において、分散 1、相関 0 でデータ数 50×5 の時とデータ数 100×5 の時、k を変化させる。

5 群の時も 3 群の時と同様、1 つの群内のデータ数が少ないと、k の値は小さい方が正解率が高くなることが分かった。3 群の時も 5 群の時もほぼ同じ場所で同じ形になっているので、データの総数ではなく、1 つの群の数に k の値は依存していると考えられるが、若干小さな値の方がよくなり、有効な範囲も狭くなるようである。今回、5 群の時は各群のデータ数が 50 個なので、5 群のシミュレーションをする時は、$k = 3$ を用いた。

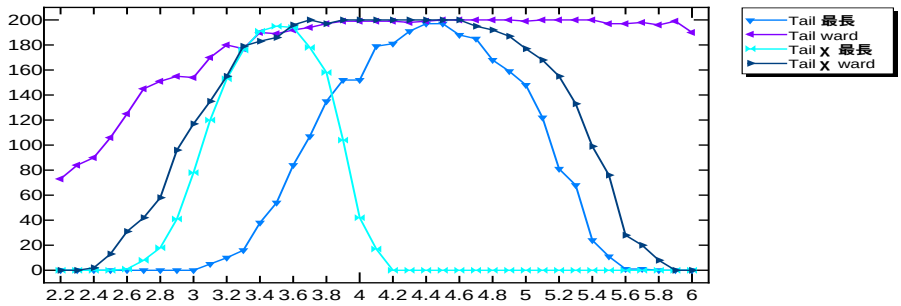

図 12: 3 群:データ数 50 × 3 の時

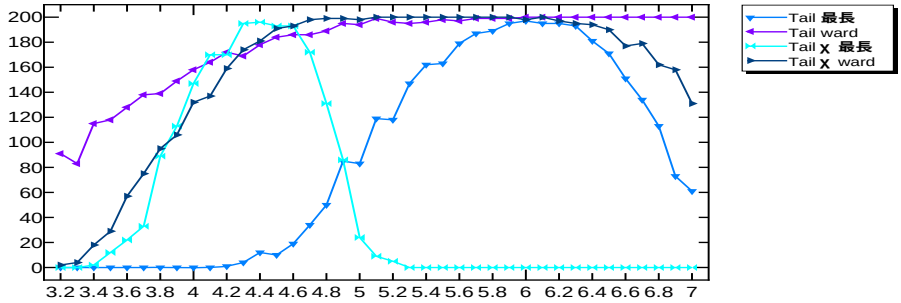

図 13: 3 群:データ数 100 × 3 の時

9.4.3 他の分布の影響

t 分布、対数正規分布に従う場合についても k を変更させてみたが、正規分布に従う時と同じ結果になり、k の値は分布よりも群の数に依存していると言える。カイ 2 乗分布を用いると、最長距離法も ward 法も正解率が高い k の範囲が狭くなっている。カイ 2 乗の ward 法の方は、ward 法のピークの間に収まる形になっているがカイ 2 乗分布を用いる最長距離法は、完全に Tail の最長距離法の k のピークがずれてしまう結果となった。

10 考察

10.1 それぞれの方法における特徴

10.1.1 JD(最長距離) 法

どの場合もすごく良い結果が出たわけではなかった。また、最長距離法は JD 法よりも Tail 法の方が合っているのかもしれない。

10.1.2 JD(ward) 法

JD の k-means 法には劣るが、たいていの場合で高い正解率を得られる。縦長なデータに弱く、まとまりのある場合であれば何群でも何分布でも、あまり影響を受けずに良い結果

が得られる。ward 法自体は、様々なクラスター数決定法にも向いていると感じた。

10.1.3 JD(k-means) 法

JD の ward 法と同じく、かなり安定して高い正解率で最適なクラスターを求めることができる。JD 法で試した中では、1 番良い結果を出した。データの分布が偏っていたり、相関や分散が少し強めな場合は正解率を下げてしまうが、それでもデータが明確に分かれている場合などは、しっかりあてる事ができる。

10.1.4 x-means 法

全体としては、あまりよい成績は得られなかった。x-means 法は、目で見て明確に判断できるようなデータは分けることができるが、そうでない場合は、信頼するのに不安が残る。今回行ったシミュレーションで用いた方法の中では、唯一単独の方法 (他は、クラスターの計算方法と停止規則を組み合わせている) でクラスター数を決定でき、クラスター数自動決定法としても確立しているので、便利ではある。

10.1.5 Tail(最長距離) 法

ほとんど良い結果が得られなかった Tail の最長距離法だが、唯一、他の方法が苦手としたパターン 2 で良い結果を得られた。k の高い正解率が得られる範囲もかなり狭く、k の値に結果が大きく左右され、クラスター数未知の場合に使うには不安要素が多い。

10.1.6 Tail(ward) 法

カイ 2 乗にすると、少しだけだが正解率が高くなる。どんな時でもかなり高い正解率が出せる。

10.1.7 Tail(最長距離-カイ 2 乗) 法

各群の分散やデータ数が全て違う場合に、圧倒的強さを見せた。実際には、分布や 1 群の個数が分からないが、k の値を外さなければ、かなりよい精度で現実の場面で使えるかもしれない。また、カイ 2 乗分布を用いる前は、ほとんど機能しない状況だったので、カイ 2 乗分布を用いたことでかなり改善されたと言える。

10.1.8 Tail(ward-カイ 2 乗) 法

Tail の ward 法の場合と同じく、また少し改善され、全体としてかなり良い結果が得られた。他の方法でうまくいかなかったパターン 2 の時にも、通用する。

10.2 k の値のとり方

Mojena[10] では、k の値を 2 から 4 で実験していた。その時の各群はデータ数 30 であったため、そうしたものと思われる。本研究の結果からも、k の値は各群のデータ数に依存していると言える。実際の解析では、各群のデータ数は分からないので、データの総数と、解析者の予想、希望するクラスター数を想定し、k の値をある程度絞ってから使えば、Tail 法の精度が上がるだろう。今回私の行ったシミュレーションでは、一環して $k = 3$ と $k = 4.5$ を用いたが、最長距離法か ward 法かでも、最適な k の値が変わってくる。各群のデータ数が 30 から 50 前後の場合、$k = 3$ 程度がちょうどよく、各群のデータ数が 100 前後になると、$k = 4.5$ 程度がちょうどよくなってくる。データ数が多くなるに従い、k の値も大きくする必要があることが分かった。しかし、データ数と k の関数関係を特定するまでは至らなかった。もっと多くのシミュレーションで詰める必要があるだろう。

データ数が全く結果が予測できないような場合は、正解率をとる範囲がかなり広い ward 法が有利である。実際、カイ 2 乗分布を用いた場合とほとんど同じ程度の精度が得られる。しかし、群の数が 5 群になると範囲は狭くなるので群の数が増えた場合にはこの有利さも限定的なものとなる。

10.3 クラスター数自動決定法を実際に用いるためには

今回、8 種類の方法において、クラスター数自動決定法を用いてシミュレーションしてきた。それぞれ得意な部分や不得意な部分があるが、実際の解析に用いるほとんどの場合、当然ながら解析対象のデータの特徴は分からない。よって、クラスター数自動決定法として用いやすく実用可能と言える方法は、やはりどんな場合にもある程度の対応ができるものであると考える。そうなった場合、JD の k-means 法や Tail の ward 法、Tail のカイ 2 乗分布を用いた ward 法があげられる。また、うまく k の値がとれるようであれば、各群の分散やデータ数がばらばらでも、高い正解率が得られる Tail のカイ 2 乗分布を用いた最長距離法が実用的であると言える。JD の k-means 法や ward 法などで、ある程度の各群のデータ数やクラスター数を予測してから、k の値を考えてもう一度 Tail のカイ 2 乗分布を用いた最長距離法で試すと、最適なクラスター数を求められる可能性が高くなるだろう。

11 おわりに

クラスター分析は、統計解析手法の中でも基本的な方法であり、また広く用いられている方法であるにも関わらず、クラスター数を自動的に決定させる方法は全く世間には知られていないのが現実であった。しかし、今回研究を進めて行く中で、統計の専門家以外にもクラスター数自動決定法を必要としている分野が沢山あり、同時に多くのクラスター数決定法が存在していることを知った。

本研究では、クラスター数決定法が実際のクラスターの群数を当てることを善し悪しの基準にしてきたが、実際のデータに対する当てはまりに関してもっと言及できればよかった。しかし、Tail 法の停止規則を求める時において、データをカイ 2 乗分布に従って変換することにより、正しいクラスター数を求める事において、かなり良い成績を残すことがで

きるようになり、従来の方法を改善することができた。ただそれらの方法も、最終的には、用いる k の値が正しければ良い結果を得る、という形になった。k の値に関しては、分けられたクラスター各群のデータ数に依存することが分かり、その動きも確認できたが、実際の分析でそれは分からないので、予測の範囲を超えることはできない。

今回、幾つかのクラスター数自動決定法を色々な方法で実際に試してみて感じたのは、シミュレーションでは良い結果を残せてはいても、それ単独でクラスター数を決定するには不安が残るという事である。様々なクラスター数自動決定法があり、それぞれに特徴があることが分かったからこそ、1 つの方法で決めようとするのではなくそれらを組み合わせることで、かなりの確率で正しいクラスター数を決定できるようになるのではないだろうかと感じた。

参考文献

[1] Hardy, A. (1996): On the number of clusters, *computational Statistics and Data Analysis*, **23**, 83-96.

[2] 石岡 恒憲 (2006): x-means 法改良の一提案 —k-means 法の逐次繰り返しとクラスターの再併合—, 『計算機統計学』, **18**(1), 3-13.

[3] 石岡 恒憲 (2006): x-means 法のソースコード

`http://www.rd.dnc.ac.jp/~tunenori/xmeans.html`(日本語)
`http://www.rd.dnc.ac.jp~/tunenori/xmeans_e.html`(English)

[4] Jain, A.K. and Dubes, R.C. (1988): *Algorithms for clustering data*, Englewood Cliffs, NJ：Prentice-Hall.

[5] Jolion, J.M., Meer, P. and Bataouche, S. (1991): Robust clustering with applications in computer vision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(8), 791-802.

[6] 神嶌 敏弘 (2003): データマイニング分野のクラスタリング手法 (1) —クラスタリングを使ってみよう！—, 『人工知能学会誌』, **18**(1), 59-65.

[7] 菅 民郎 (1993): 『多変量解析の実践（下）』, 現代数学社.

[8] Krishnapuram, R. and Freg, C.P. (1992): Fiting an unknown number of lines and planes to image data through compatible cluster merging, *Pattern recognition*, **25**, 385-400.

[9] Marriot, F.H.C. (1971): Practical problems in a method of cluster analysis, *Biometrics*, **27**, 501-514.

[10] Mojena, R. (1977): Hierarchical grouping methods and stopping rules: an evaluation, *The Computer Journal*, **20**, 359-363.

[11] Moore, M. (1984): On the estimation of a convex set, *The Ann.Statist.*, **12**, 1090-1099.

[12] Nasraoui, O., Leon E., and Krishnapuram, R. (2005): Unsupervised Niche Clustering: Discovering an Unknown Number of Clusters in Noisy Data Sets. Chapter 8, pp.157-188, in *Evolutionary Computing in Data Mining*, A.Ghosh L.C.Jain, Eds, Springer Verlag.

[13] Ngo, C.W., Pong, T.C. and Zhang H.J. (2002): On clustering and retrieval of video shots through temporal slices analysis, *IEEE Trans. Mlt.*, **4**(4), 446-458.

[14] Pelleg, D. and Moore, A. (2000): X-means: Extending k-means with efficient estimation of the number of clusters, `http://www-2.cs.cmu.edu/~dpelleg/download/xmeans.pdf` .

[15] 志津 綾香 (2009): クラスター数決定法についての研究, 2008 年度南山大学数理情報学部数理科学科卒業論文.

[16] 田中 豊・脇本 和昌 (1983): 『多変量統計解析法』, 現代数学社.

[17] 渡辺 洋・南風原 朝和・大塚 雄作・石塚 智一・山田 文康・藤森 進・前川 眞一 (1988): 『心理・教育のための多変量解析法入門-基礎編』, 福村出版.

[18] Wolfe, J.H. (1970): Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Res.*, **5**, 329-350.